View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by DSpace at VU

Journal of Official Statistics, Vol. 22, No. 3, 2006, pp. 419–451

Methods of Behavior Coding of Survey Interviews

Yfke P. Ongena[1,2] *and Wil Dijkstra*[3,4]

Comparing 48 different coding schemes, we attempt to give an exhaustive overview of all methods of behavior coding of survey interviews. Coding can take place at the level of the utterance, of the exchange or of the whole question–answer sequence. If the sequence is used as a coding unit, the complexity of the coding scheme will be low but so will the amount of information in the data. If the utterance is used as a coding unit, it is possible to apply full coding (i.e., all utterances are coded) or selective coding (only relevant utterances are coded). Full coding of utterances with preservation of sequence information is by far the most labor-intensive but also the most informative, as a lot of information can be derived from sequence analyses. In that case it is advisable to use a multivariate coding scheme. More simple coding schemes are advised when frequency analyses are applied.

Key words: Survey interviewing; question–answer sequence; interviewer monitoring; pre-testing methods; interaction analysis.

1. Introduction

The importance of studying the interviewing process has gained more and more recognition over the past 30 years. Cannell, Fowler, and Marquis (1968) concluded that within the interview itself, particularly in the behavior of the participants, we can find the most important causes of good and poor survey responses. Although the first studies were primarily directed towards the behavior of the interviewer in order to detect bad interviewer performance, it soon became apparent that the behavior of the respondent is equally important in understanding the question-answer process. The relation between validity of responses and the occurrence of problematic behaviors in interviews has been demonstrated in several studies (e.g., Belli and Lepkowksi 1996; Dijkstra and Ongena 2002; Dykema, Lepkowski, and Blixt 1997).

A twofold answer to the question "Why study interaction in survey interviews?" was provided by Van der Zouwen (2002). His first answer refers to the revealing of (either positive or negative) effects of the interaction itself on the responses obtained, i.e., using the method as a diagnostic instrument. The second answer refers to the revealing of

[1] University of Nebraska-Lincoln, Survey Research and Methodology Program, 200 North 11th Street, Lincoln NE 68588-0241, U.S.A Email: YOngena2@unl.edu
[2] Department of Social Research Methodology, Vrije Universiteit, Amsterdam, The Netherlands.
[3] Department of Social Research Methodology, Vrije Universiteit, Email: W.Dijkstra@fsw.vu.nl
[4] NIAS, Wassenaar, The Netherlands. An earlier version of this article was presented at the Workshop "Methods for Studying Interaction", University of Wisconsin-Madison, April 12-14, 2002.
Acknowledgments: The authors thank three anonymous reviewers for their comments on earlier versions of this article.

© Statistics Sweden

difficulties that interviewers and respondents themselves have in questioning and answering, i.e., using the method as a problem-solving instrument.

Behavior coding comprises a systematic coding of interviewer and/or respondent behaviors in survey interviews. The process of questioning and answering in the survey interview that it studies, takes place in so-called question-answer sequences (Q-A sequences), which comprise all utterances of interviewer and respondent that belong to a survey question.

Both the interviewer and the respondent can cause deviations from the so-called "paradigmatic" sequences. Schaeffer and Maynard (1996) introduced this term to indicate sequences that are perfect from a survey researcher's point of view. During a paradigmatic sequence (or "straightforward sequence," Sykes and Morton-Williams 1987) the interviewer poses the question as scripted, the respondent gives an adequately formatted answer that is assumed to be appropriate, and the interviewer may neutrally acknowledge this answer.

In a broad sense, behavior coding is intended to discover departures from the paradigmatic sequence, and to discover how these departures relate to data quality on the one hand, and characteristics of interviewer, respondent, or questionnaire design on the other. Paradigmatic sequences usually make up the largest part of Q-A sequences in an interview, but may vary from for example 35% to 95% of the Q-A sequences for different questions within the same survey (Van der Zouwen and Dijkstra 1998).

In 1968, Cannell, Fowler, and Marquis devised the first, fairly simple scheme to code behavior in the standardized survey interview. Next coding schemes generally became more and more sophisticated as well as more complex, as with each subsequent coding scheme and its application to actual data, more and more became known about the interaction between interviewer and respondent. In addition, the development of more sophisticated coding schemes was stimulated because technical devices became available. Especially the availability of the tape recorder may explain the increase in the number of codes that were included in the coding scheme. The scheme of Cannell, Fowler, and Marquis (1968), including only twelve different codes, did not rely on the availability of tape recorders. In a subsequent study, Marquis and Cannell (1969) did use tape recordings, and described a far more detailed coding scheme, consisting of 47 different codes.

The increase in the number of codes that could be included in coding schemes was even more stimulated by a second technical device that could be used for behavior coding. This device was the computer. A program like the Sequence Viewer program (Dijkstra 1999 2002) enabled the coder to quickly and reliably enter a lot of different codes, and the coding could also be carried out semi-automatically, based on the transcripts. The text analysis options in this program enable automatic coding of all paradigmatic Q-A sequences. However, the increased feasibility of entering large amounts of data was not the only benefit of the use of computers. The possibility of *analyzing* a large number of codes and large data sets was another major advantage of using computers. Because of that capacity, it became worthwhile to invest in the time-consuming process of transcribing and coding interviews in a detailed way. For example, Loosveldt (1985) describes that for the analysis of the 11,331 actions that were coded, special programs were written. The Sequence Viewer program also allows researchers to perform a large number of different, more and more sophisticated analyses (Dijkstra 2002).

The number of different categories included is probably the most obvious difference between coding schemes. The number of categories varies from two values (Edwards et al. 2002) to around two hundred different code combinations in an average dataset (Dijkstra 1999).

It is beyond the scope of this article to give a full account of all codes used in the 48 coding schemes that were studied, but we will discuss some common distinctions. We found 134 different categories for interviewer behavior, 78 different categories for respondent behavior, and 14 different categories for behavior of third parties (see Table 1 for examples typical behavioral codes).

Cannell and Oksenberg (1988) indicate that the kinds of code categories that are included in a coding scheme depends upon the research objective. However, this appears to be only partially true; irrespective of the focus of the scheme, most schemes include codes for interviewer's question reading.

For behavior coding as a proper diagnostic tool, it is important that all relevant behaviors are included in the coding schemes. It may not always be possible to determine in advance what those relevant behaviors are. Therefore the development of a behavior coding scheme can be considered an iterative process.

Table 1. Most common codes included in coding schemes and average reported frequency of occurrence in Q-A sequences

Interviewer behavior codes	Number of coding schemes	Range in percentage of occurrence	Respondent behavior codes	Number of coding schemes	Range in percentage of occurrence
Question read exactly as scripted	26	28–97%	Adequate answer	25	75–95%
Question read with minor change	21	1–32%	Inadequate answer	21	2–27%
Question read with major change	35	0–25%	Don't know answer	17	1–6%
Question skipped/not verified	16	0–22%	Refusal to answer	21	0–1%
Non-directive probe in interviewer's words	23	5–80%	Request for clarification	18	0–23%
Suggestive probe	15	0–33%	Interruption	18	0–36%
			Qualified answer	14	2–20%

Note: The codes listed are used in at least 12 (i.e., 25%) of the 48 coding schemes evaluated in this article. The range in percentage of occurrence applies to occurrence of the behavior in Q-A sequences as reported in the studies that used the code.

As Table 2 shows, behavior coding is typically related to variables in the data collection procedures (i.e., question wording, interviewer styles etc), and can be implemented in different phases of survey data collection. Results of behavior coding implemented prior to or during actual data collection can be used to adapt data collection procedures. Behavior coding data can also be used as dependent variables in experiments (e.g., comparing question wordings or differently trained interviewers). They can also be used as independent variables in studies that aim to detect relations between problematic behaviors and the validity and reliability of scores obtained (Belli and Lepkowski 1996; Dykema, Lepkowski, and Blixt 1997; Dijkstra and Ongena 2002).

In this article an exhaustive overview is given of all applications of behavior coding, comparing characteristics of 48 coding schemes,[4] presented in manuals, conference proceedings, articles etc. Advantages and disadvantages of different strategies and procedures will be given. Finally we give recommendations about the types of coding schemes and procedures that are most appropriate in specific situations.

2. Coding Strategies

Some fundamental decisions in the design of a coding scheme have consequences for the applicability of the scheme. These decisions concern the unit of coding, whether full or selective coding is applied, and whether and how sequence information will be preserved.

2.1. Units of Coding

Behavior coding most typically occurs at one of four levels: (1) individual utterances, (2) exchange, (3) Q-A sequences or (4) entire interviews. These levels are described below.

2.1.1. Coding at the Utterance Level

A strategy that is especially useful in interaction analysis is coding at the level of the utterance. Each utterance can get one code, but not more than one code. It is not possible to code utterances that did not take place, e.g., the absence of an adequate answer. However, if full coding is applied (see Section 2.1.5), and/or sequence information is preserved it is possible to infer the absence of certain behaviors from the coded utterances within a Q-A sequence.

To code the utterances of a Q-A sequence, the sequence should be separated into meaningful parts. The turn is too rough as a segmentation procedure, because it may consist of multiple "turn-constructional units" (TCU's), utterances that can be considered fully informational units. They are constructed in such a way that other speakers are able to determine when and whether the TCU is complete (Sacks, Schegloff, and Jefferson 1974). When coders try to determine the appropriate codes, most problems occur as soon as utterances are not adequately segmented into separate TCU's. Multiple types of behaviors can be performed within a turn. As a result, multiple codes may be applicable to one turn, which creates a problem for the coder.

[4] In this comparison of coding schemes only first published articles concerning coding schemes are included. Coding schemes of the same author(s) that underwent important changes (either in the codes included or in the coding procedures) are treated as separate cases.

Table 2. Possible implementations of behavior coding

Goal	Phase of study
Pretest questionnaire, interview mode etc.	*Prior* to actual data collection
Monitor interviewers	*During* actual data collection
Evaluate data quality, functioning of interviewers and respondents, effectiveness of revisions, explain biases in response distributions	*After* actual data collection
Explore causes and effects of behaviors	*After* actual data collection
Check experimental manipulations	*After* experimentally manipulated data collection
Use behavior coding as a *dependent* variable	*After* experimentally manipulated data collection

In interviewer scripts multi-unit turns are often present (i.e., interviewers have to read introductions, instructions, response alternatives, specifications and questions; see Houtkoop-Steenstra 2000). Respondents may also perform multiple behaviors in one turn. Therefore, it is important that the utterances in Q-A sequences are carefully segmented into TCU's. According to pragmatic completeness, a TCU is complete when the utterance is recognizable as an independent informative and functional unit. Pragmatic completeness is assessed by means of sequence reasoning, i.e., the sequential position of an utterance as part of sequences that are functionally related (Mazeland 2003). Segmenting the utterances consists of judging the pragmatic completeness of utterances, whereas coding the utterances consists of applying a pragmatic description to each one.

2.1.2. Coding at the Exchange Level

It is possible to code at a level that is intermediate between the utterance and the Q-A sequence levels; this intermediate level is often referred to as the exchange level. An exchange can be considered an adjacency pair of a question and an answer. Typically, the first two exchanges are coded, i.e., (1) the exchange of initial question reading and an initial response, and (2) the exchange of a prompt by the interviewer and a possible second answer by the respondent. The coder must ignore insignificant behaviors that may occur in between (e.g., neutral acknowledgement tokens, silences, laughter) and ignore anything after the second answer. Morton-Williams (1979) was the first to use this kind of coding. Such a coding strategy is selective with respect to the part of the Q-A sequence that is coded, but it still enables preservation of sequential information, which is not possible in the case of coding at the Q-A sequence level.

2.1.3. Coding at the Q-A Sequence Level

Assigning a code to the whole Q-A sequence may involve judging whether or not a specific type of behavior takes place in the Q-A sequence, or whether or not the Q-A sequence is paradigmatic or problematic. The division of units to be coded is in this case more straightforward: a Q-A sequence starts as soon as the interviewer starts reading a question, and ends as soon as the next question is posed. However, it is of course possible

that, whereas the interviewer has posed a next question, the respondent elaborates his or her answer to the previous question. Such behaviors may be easily overlooked, or assigned to the wrong Q-A sequence, especially when coding does not take place from transcripts (see Section 3.1).

As compared to coding at the utterance or exchange level, coding at the Q-A sequence level is more sensitive to errors of omission. According to Cannell, Lawson, and Hausser (1975), disagreements in coding of entire Q-A sequences often do not occur in respect of the choice of a particular code to be used for a behavior, but rather in respect of whether or not a particular behavior should be coded at all.

2.1.4. Coding at the Interview Level

A final unit is the whole interview, e.g., if the whole interview is assigned some evaluative code. Carton (1999), for example, added codes to characterize the whole interview with respect to specific interviewer behaviors such as giving instructions, asking questions and probing, and general evaluations such as the orientation towards the respondent and the atmosphere during the interview. In the comparison of behavior coding schemes we did not include schemes that only use coding at the level of the interview (e.g., Brick et al. 1997a; Mathiowetz 1999)

2.1.5. Full or Selective Coding

A fundamental difference between coding schemes is that coding can be applied to all utterances ("full coding") or to a selection of utterances or behaviors that are considered important or relevant for the specific research question ("selective coding"). Selective coding schemes are essentially developed from a practical point of view: it is determined in advance what behaviors are diagnostic of problems that the researcher wishes to detect. For example, if one studies general interviewer performance, only interviewer behaviors are coded.

A full coding scheme is often used when the researcher's goal is to explore the interaction. With full coding data it is possible to reconstruct more or less what occurred in an interview. Full coding must take place at the utterance level, as it requires assigning a relevant category to each utterance, whereas selective coding may take place at the QA-sequence level or at the utterance or exchange level. In the latter two cases, it is possible to preserve sequential information at the exchange level. For example, in Cannell, Lawson and Hausser's (1975) coding scheme only interviewer behaviors were coded (therefore constituting a selective coding scheme at the utterance level). Nevertheless, they instructed the coders to code in the order of occurrence, and all respondent utterances in between the interviewer's utterances were represented by vertical lines.

The combination of the three levels of coding and application of full or selective coding yields six possibilities, of which only four are relevant, because full coding can only take place at the utterance level. Hence we can distinguish four coding strategies; full coding of utterances, selective coding of utterances, coding at the exchange level and coding of whole Q-A sequences. These strategies have different consequences for the possibility of preservation of sequential information, as shown in Table 3.

In Table 4 advantages and disadvantages of three coding strategies are shown. Coding at the Q-A sequence level makes quick results possible, without the use of specialized

Table 3. Overview of coding strategies and possibilities of preserving sequential information

Strategy	Unit of coding	Sequential information applicable
Full coding	Utterance	++
Selective coding	Utterance	+
	Exchange	+
	Q-A sequence	−

software. For instance, coders may only have to note inadequate readings of questions or requests for clarification from respondents.

Full coding is by far the most tedious kind of coding. In order to apply full coding, it is important to have software available that facilitates the transcribing, coding and analyzing of the data. Without such software, full coding with sequential information is hardly feasible.

As Smit (1995) argues, it is important that the number of codes included in a coding scheme is manageable; with too detailed coding schemes it will often be problematic to employ clear methods of analysis. Moreover, with a complex coding scheme the coding process will be more error-prone and time-consuming. For full coding a detailed and consequently complex coding scheme is necessary to meaningfully characterize all the various behaviors that can occur during an interview. However, several options are available to enhance the simplicity of the scheme (see Section 3.4).

Whole Q-A sequences can easily be coded according to the *absence* of relevant behavior. In the case of full coding, absence of behavior may be inferred from analysis of complete Q-A sequences.

The *amount of information* will usually be lowest in case of coding at the Q-A sequence level, hence potentially important behavior may easily be overlooked. Most information, also about the *sequence* of behaviors, is available in the case of full coding; it provides a researcher with information about any *deviation from a paradigmatic sequence*. In the case of coding at the Q-A sequence level, it is possible to include codes that evaluate the Q-A sequence as a whole. In case of selective coding of utterances or exchanges, it is

Table 4. Overview of advantages and disadvantages of coding strategy

	Selective coding: whole Q-A sequence	Selective coding: utterances or exchanges	Full coding: utterances
Quick results	Yes	Moderate	No
Practical feasibility	Software not necessary	Software may be helpful	Hardly feasible without software
Complexity	Low	Low	High
Absent behavior	Possible	Difficult	Can be inferred
Amount of information	Low	Medium	High
Sequence information	Not available	Possible	Available at no extra cost
Identification of paradigmatic sequence	Possible	Difficult	Always available

difficult to obtain information on all deviations from paradigmatic sequences. In all cases of selective coding, it is possible that deviations that are not coded are more indicative of problems than the coded ones.

2.2. *Type of Analysis*

Two main types of quantitative analysis of behavior coding data can be distinguished, i.e., frequency analysis and sequence analysis. Furthermore, quantitative analyses may be supported by qualitative analyses of the actual interactions, provided that transcripts are available.

2.2.1. Frequency Analysis

Frequency analysis essentially concerns counting the occurrence of particular types of interviewer and respondent behavior. The frequency of occurrence of specific behaviors may be related to other factors, like interviewer or question characteristics, or response distributions. For example, Edwards et al. (2004) compared frequencies of interviewer and respondent behaviors across interviews of the same questionnaire in different languages. One of the findings was that respondents appeared to behave differently when they were being interviewed in their first language (i.e., interrupting the interviewer and making extraneous comments more often) than in a second language.

Furthermore, frequency analysis can be used in experimental designs that compare manipulations of data collection procedures in survey interviews. For example, one can establish the effects of different question wordings on the occurrence of inadequate answers.

Frequency analyses can be supplemented with analyses of variance or log-linear analyses at the Q-A sequence level (i.e., comparing question, interviewer or respondent variables with average number or odds ratios of problematic behaviors occurring in the Q-A sequences).

2.2.2. Sequence Analysis

Sequence analysis allows studying dependencies between different types of behavior, in particular the relation between subsequent interviewer and respondent behaviors. In the case of selective coding schemes, sequence analysis is rather limited; it is possible to distinguish initial from secondary responses, and initial question asking from follow-up probing, but not for example what kind of nonproblematic behaviors may have occurred in between questions and answers.

In order to be able to interpret the results of sequence analysis correctly, it is important that the assignment of codes is independent of codes that precede or follow the behavior to be coded. In some cases it is hardly avoidable that coding a particular behavior depends on previous utterances. A code for "interviewer repeats respondent's answer" is likely to be preceded by an answer from the respondent. Therefore it is hardly possible not to take the preceding utterance into account. Nonetheless, assigning a particular code should never depend on subsequent behavior, to prevent relations between behavior and subsequent behavior from being artificial.

Data that are generated through full coding schemes enable analyses by means of a *tree* representation of the structure of interviewer–respondent interaction. Brenner (1982) was the first researcher to present such a tree analysis. A tree may represent the consequences of a particular action of either interviewer or respondent. From other analyses it is possible to analyze the causes of particular actions of interviewer or respondent. For example, with the lag-sequential analysis that Smit (1995) describes, it is possible to determine which parts of subsequent behaviors in a Q-A sequence occur below or above chance.

How sequence analyses may also be helpful to describe interactional processes can be illustrated by means of findings of Dijkstra and Ongena (2002). They found that a mismatch answer (i.e., an answer that is not formatted according to the prescribed alternatives) is not only the most frequently occurring respondent problem; it is also an important cause of problematic interviewer deviations. Furthermore, they showed that when interviewers repeat the response alternatives after such a mismatch answer, they more often immediately obtain an adequate answer than when they repeat the entire question.

2.2.3. Supplementary Analyses

Behavior coding studies concerning the frequency of occurrence of behaviors very often only give data from tables and do not uncover sources of problematic behaviors. It often remains unclear, even in the case of sequential analysis, how events in the interaction can have certain causes or effects, i.e., what actually happened in the interaction.

One way to learn more about this, is to use code frequencies as input for discussions with interviewers or coders (i.e., debriefing; see Oksenberg, Cannell, and Kalton 1991). Using coders for debriefing is useful because coders have no personal involvement in the interviewers and, having listened to tape recordings, have full access to relevant information of the interactions (DeMaio et al. 1993). Notes of coders are often used to diagnose the sources and the seriousness of the problems (e.g., Dykema, Lepkowski, and Blixt 1997; Schaeffer and Dykema 2004). Such notes may specify a major change in question reading, with abbreviations to indicate the nature of the change (addition, deletion or other) and the indications of the specific words that were added or deleted (Schaeffer and Dykema 2004).

However, the actual conversations on tape could be even more useful. It is quite possible that coders do not notice all interesting aspects that are worth inspecting in more detail. Furthermore transcripts can easily illustrate findings. Furthermore, other sources of information can be used, such as answer distributions, response latencies (see Draisma and Dijkstra 2004) and details of the date, time and location of the interviews.

3. Practical Considerations in Coding Procedures

The coding procedure is an important feature when it comes to the usability and reliability of a coding scheme. According to Cannell and Oksenberg (1988) it makes little difference whether the observation mode comprises face-to-face or telephone interviews, and whether live coding or coding from tape recordings is used, because the techniques for coding behavior are the same. However, they ignored the procedure of using transcripts, which is hardly to be avoided in the case of full coding, but an option in the case of selective coding.

3.1. Live Coding, Coding from Tape and Using Transcripts

Coding can be done during the interview ("live coding") or afterwards, by listening to tape-recorded interviews ("recorded coding") or by using transcripts of the tape-recorded interviews ("transcript coding"). The advantages and disadvantages of these three procedures are summarized in Table 5. The elements listed in the table may differ in importance, depending on the research question and objectives at hand.

In only six studies is some indication given of the time involved in coding interviews (including transcribing or otherwise). This ranges from a time equal to the interview, in the case of live coding, to about six times the duration of an interview, in the case of transcript coding.

The advantage of live coding is of course that data are immediately available; it is finished concurrently with the interview. Coding from tape may be more *efficient* than live coding, because coders do not have to wait for an interview to occur (DeMaio et al. 1993). Furthermore, tape coding is a relatively quick method, because no transcripts are produced. However, the additional time that is needed for producing transcripts may be regained when complex Q-A sequences are coded. In that case transcripts may help coders to see, the complete Q-A sequence. With this information it is easier to determine what code is appropriate, and in case of doubt it is possible to just read the utterances in the transcript again instead of rewinding the tape to search for the fragment.

In the case of live coding, *permission* to record the interview is of course not necessary. However, live coding in the case of personal interviews may be more *obtrusive* than coding from tape or transcripts, because a coder needs to be present during the interview.

Although live coding can be *reliable* (Esposito et al. 1992), recorded coding will always enable better quality of coding, as coders have more time to decide on the most appropriate code, and can consult code descriptions. Transcript coding in fact comprises a coding procedure in three steps (transcription, segmentation of meaningful utterances, and coding, comprising assignment of meaning to utterances). The researcher may perform separate reliability checks for the two latter tasks (see Smit 1995), or even decide to assign the different tasks to independent transcribers and coders.

Table 5. An overview of advantages and disadvantages of different coding procedures

	Live coding	Live coding with tape as backup	Recorded tape coding	Recorded transcript coding
Cost	Low	Low	High	Highest
Permission	Not needed	Needed	Needed	Needed
Obtrusive	Yes	Yes	No	No
Efficient planning	No	No	Good	Moderate
Reliability	Low	Low	Better	Better
Semi-automatic coding	No	No	No	Yes
Check of coder performance	No	Yes	Yes	Yes
Paralinguistics	Hardly	Hardly	Yes	Uncertain
Thorough analysis	No	Low	Moderate	High

Whenever coding takes place live or direct from tape, it is likely that important, meaningful behaviors are ignored. It is important that coders have useful visual documents available that enable them to compare what they hear on tape with the exact question wordings and the interviewer's recordings. Completed questionnaires or responses that are copied onto blank questionnaires may be an alternative to transcripts (Cahalan et al. 1994). However, especially complex coding schemes will require transcripts to warrant reliable coding. As Dijkstra (1999) points out, coding from transcripts can be done *semi-automatically* for utterances that occur frequently.

Tape coding enables *checks of coder performance*, but transcript coding enables more systematic checks. Determining inter-coder reliability in the case of live coding is only possible by means of having multiple coders code simultaneously. However, a live-coded interview may be taped as well, so as to be able to check samples of the coding and to (re)code or correct complex parts of the interactions. In that case some advantages of recorded and live coding are combined.

In some cases special attention must be paid to *paralinguistic* features of the utterances. A different tone and accent can for example change the meaning of an utterance. When just the written text is used for coding, errors might be made as a result of ignoring these features. It is therefore important to have sound files easily available when coding from transcripts.

Obviously, recorded coding as compared to live coding increases the options in the complexity of the coding scheme and thus makes more *thorough analysis* possible. But, as noted before, transcripts certainly will be helpful to illustrate or explain results from plain analysis of the codes. When the interview is coded from tape, it will be less likely that effort will be invested to find the fragment that illustrates a certain result.

It appears that recorded tape coding is the most popular procedure, as in 31 of the 48 schemes this procedure was followed. The difference between live coding and recorded coding is clearly illustrated by the number of codes included in coding schemes. Schemes that are designed for live coding contain between 2 and 20 codes (median: 13 codes), whereas schemes designed for recorded coding contain 2 to 174 codes (median: 22 codes). The schemes designed for recorded transcript coding contain between 15 and 199 codes (median: 30 codes).

3.2. Use of New Technologies

In line with the latest developments, interviews may be recorded as a digital sound file. In this way the computer is not only used as a device to go through a questionnaire (CATI or CAPI), but also enables "Computer Audio Recorded Interviewing" (CARI), using the computer as a "sophisticated tape recorder" (Biemer et al. 2000). Because no additional recording device such as a tape recorder is visible, recording is less obtrusive and respondents and interviewers are more likely to forget about the recording during the interview. With CARI the software instead of the interviewer controls recording, and arrangement of recording (e.g., to start concurrently with the interview or skip recording at specific sections) can be integrated with CATI/CAPI software (see Ongena, Dijkstra, and Draisma 2004).

As Shepherd and Vincent (1991) argue, when coders compare question wording with interviewer's wording "they need to review a questionnaire source document that is

identical to the document used by the interviewer" (Shepherd and Vincent 1991, p. 529). Therefore, when it comes to interviews that are computer-assisted, ideally an electronic version of the questionnaire needs to be available, e.g., to account for complex skip patterns and automatically adapted question wordings. In Shepherd and Vincent's study, the coders used the CAI program itself, in order to view the questionnaire in exactly the same way as how interviewers had it available during the interview. In the Sequence Viewer program (Dijkstra 2002), several sections on the screen are available for coders with information on the exact question wording, the response alternatives and show cards used in the interview.

3.3. Availability of Code Descriptions

In order to warrant the reliability of results it must be clear to what kind of behaviors a coder should apply certain codes. Interpretation of results will certainly be difficult if coders did not uniformly understand when to apply which code. Of course it is impossible to provide descriptions of all possible ambiguous situations. Therefore it is useful to document extraordinary situations by letting coders make notes on the ambiguities they came across in coding. The researcher can subsequently use these notes to adapt instructions for all coders.

Authors often give only an overview of the codes they used, and only indicate the code with two or three words ("adequate answer," "inappropriate probe" etc). Some authors (e.g., Cannell, Lawson, and Hausser 1975; Prüfer and Rexroth 1985; Snijkers 2002) present their codes more clearly in that they give a short description (e.g., "makes up in own words a probe (query) which is nondirective").

Brenner (1982) is one of the authors who present their codes the clearest, by not only describing them but also giving fragments of Q-A sequences to illustrate them. Dijkstra (1999) uses the same strategy with clear examples, which are essential to explain the multivariate coding scheme (see Section 3.4).

3.4. Organization of the Coding Scheme

In the case of a large number of codes, it is important that the coder is able to manage this large number, to quickly choose the right code. This management is obviously improved when codes are well organized, for example by means of grouping them in similar categories of behavior. These categories may also be a means to reduce the number of codes, when for some analyses the different codes within a category are treated as one category. Cannell, Lawson, and Hausser (1975), for example, grouped their codes into limited sets of interviewer activities, such as "posing questions," "probing and clarifying," and "other behavior." These sets were each arranged in two groups of correct and incorrect behaviors. The codes consist of two digits, with the first digit indicating the code category (e.g., "correct question reading") and the second a further specification (e.g., "reading the question exactly as worded"). It is therefore possible to use a reduced version of the coding scheme, using only the first digit.

In Dijkstra's (1999) *multivariate* coding scheme the behaviors of the interviewer and respondent are coded on a number of different coding variables. The coder, accordingly, needs to make several decisions (i.e., for each variable) when coding one utterance.

Instead of making one decision concerning the choice between up to hundred different codes, as in the schemes of Blair (1978) and Prüfer and Rexroth (1985), the coder makes the same decision in multiple small steps. Using this procedure, the coders need to memorize only a relatively small number of codes, whereas the combination of the code variables may result in a very large number of different codes. A multivariate scheme may be more reliable than a univariate one, because when coders choose the wrong code values on one variable, the other variables may be correctly coded (Dijkstra 2002). Loosveldt (1985) used a similar strategy, and also Mathiowetz and Cannell's (1980) and Blair's (1978) coding schemes can be considered multivariate.

3.5. The Coders

The validity and reliability of the results obtained with the coding scheme depends on the persons who did the coding. As experimental research in social psychology has shown, observers may draw on specific theories when assigning meaning to behavior. For example, observers are more likely to draw on what they know about the actor's character in explaining behavior than when they explain their own behavior (for a review of experimental studies, see Watson 1982). Therefore coders need to be trained, especially in case of complex coding schemes.

Coders may be biased by the researcher's expectations and make inferences based upon these expectations. Bakeman and Gottman (1997) state that it is important not to inform coders about hypotheses of a behavior coding study. In addition, they point out that not only inter-coder reliability is important, but also intra-coder reliability. Especially in case of complex coding schemes and when the coding process takes a long time, the coding may lose consistency. Moreover, it can hardly be avoided that coders develop their own expectancies during coding. A useful check is to compare codes assigned during the first half of the coding work with those assigned during the second half.

3.5.1. Researchers

Some researchers (Brenner 1982; Loosveldt 1985; Van der Zouwen and Smit 2004) did the coding themselves, almost turning behavior coding into some kind of expert review. Apparently they only trust themselves in grasping the subtleties of such coding schemes. As Brenner (1982) states: "it proved impossible to find people who were willing, against payment, to code the tapes to a sufficiently high standard" (Brenner 1982, p. 143).

A disadvantage of this strategy is that not only coding may be biased by the researchers' hypotheses about the outcomes, but also the coding scheme may be less appropriate to be used reliably by other researchers. Therefore, reliability scores of studies with researchers doing the coding themselves should be interpreted with care.

3.5.2. Field Staff

A second possibility is to use field staff: either experienced interviewers who did not participate in the survey being coded, or supervisors, "control staff," "researchers" or "methodologists" as coders. An advantage of using this group is that these persons are familiar (or ought to be familiar) with interviewing conventions.

In the studies of Burgess and Patton (1993) and Snijkers (2002), the interviewers participating in the survey did the coding (of respondent behavior) themselves during the interview (using 5 and 7 different codes respectively). According to Burgess and Patton, coding could be applied easily, as "proven" by perceptible delays in the interviews of "only" 2–3 seconds for each code to be entered, which "added perhaps 10 seconds on average to the length of the interviews, which averaged over 30 minutes" (Burgess and Patton 1993, p. 396). In Burgess and Patton's (1993) study less than 3% of the Q-A sequences received a code. However, it is very unlikely that the target behaviors (i.e., respondent asks for repetition or clarification, interrupts interviewer, asks the time left for the interview, or seems uncomfortable) occurred in only 3% of the Q-A sequences. Therefore this clearly illustrates that an interviewer is not capable of capturing all occurrences of behaviors that need to be coded. Moreover, the fact that interviewers are coding the respondent's behavior may itself influence the interaction, as suggested by a side-effect that Snijkers (2002) found: it appeared to make interviewers more alert to problems with questions.

3.5.3. Trained Coders

A third group of coders are specially trained coders, who do not necessarily have interviewing experience. Unlike when it comes to using interviewers as coders, these coders should also be trained with respect to interviewing conventions.

Coders may be provided with oral descriptions of the coding scheme and its application, followed by practical sessions with feedback from the researcher (Sykes and Collins 1992), or a manual with exercises (Dijkstra, Van der Veen, and Van der Zouwen 1985). The length of training may vary from one to two hours individual training (Blair 1978) to 45 hours (Oksenberg, Cannell, and Blixt 1996). Training of coders may also take place with a simultaneous further development of the coding scheme (Belli et al. 2004).

4. Reliability of the Coding Scheme

In 23 studies reliability scores are presented. Unfortunately, researchers do not use the same methods of determining reliability. Moreover they do not all present their methods clearly; therefore we can often only guess how reliability scores were produced.

Reliability checks should be done with samples of multiple interviewers and respondents. It is better to double-code random parts of multiple interviews than to double-code one or more complete interviews, because both interviewer and respondent styles may greatly differ, and more differences between interviews will be found than within one interview (Cannell, Lawson, and Hausser 1975).

Generally, the best way to test reliability is to test it at the same level as the level that was used for assigning codes. The more general the level, the less informative reliability scores are. For example, when codes are applied at the Q-A sequence level we only know if coders agree that a certain behavior occurred in a Q-A sequence; we do not know whether or not coders based this decision on the same utterance. It is perfectly possible that multiple instances of the same behavior take place within the same Q-A sequences. Therefore, reliability scores at the Q-A sequence level are generally overestimated.

Agreement scores at the utterance level can be divided into two different types: agreement upon what should be considered a separate utterance and agreement upon the individual codes (Smit 1995). However, in most behavior coding studies reliability of these two types of agreement is not established.

Researchers are not uniform in their use of statistics for reliability testing (i.e., Kappa statistics, Pearson correlations or simple percentages). Percentages of agreement are computed by dividing the number of units with the same code by the total number of units coded. When the coding scheme contains only few different codes, the probability of chance agreement is very high. In the Kappa statistic the probability of chance agreement is incorporated.

In a number of cases the authors give detailed reliability information, e.g., separate reliability scores for interviewer and respondent behaviors, or even for each separate code category, which in some specific cases is quite low (cf. Blair 1978; Oksenberg, Cannell, and Kalton 1991; Belli et al. 2004; Edwards et al. 2004). A low reliability score may be the result not only of ambiguity between two or more specific code categories, but also of the absence of adequate code descriptions, inadequately skilled coders, or an inappropriate coding procedure.

The negative relationship between code complexity and accuracy is often demonstrated (see e.g., Dorsey et al. 1986). Intuitively it makes sense that accuracy and inter-observer agreement are higher when the coding task is simpler. However, the correlation between the number of codes included (as a measure of coding scheme complexity) and the overall reliability score of Kappa values appeared to be positive but nonsignificant ($r = .166$, $p > 0.05$, $n = 16$). Kruskall-Wallis tests showed that neither differences in reliability scores were related to the strategy (full, selective, or sequential, $\chi^2 = 3.23$, $\text{df} = 2$, $p > 0.05$, $n = 16$), the procedure (transcript coding, live, or recorded coding, $\chi^2 = 3.55$, $\text{df} = 2$, $p > 0.05$, $n = 16$) or the kind of coders (researchers, field staff or trained coders, $\chi^2 = 2.46, \text{df} = 2$, $p > 0.05$, $n = 16$) used.

5. Focus of the Coding Scheme

Bakeman and Gottman (1997) state that creating a coding scheme is theoretically based, because the coding scheme represents a hypothesis. The scheme contains behaviors and distinctions that a researcher considers important. Therefore, they argue that researchers can only rarely use the coding schemes of others. A different research question indicates a different coding scheme, and this would imply that comparing coding schemes developed for different research questions is not useful.

However, this might be less true for coding schemes designed to describe the behavior in standardized survey interviews. As Table 1 already indicated, quite a large degree of overlap can be found in the codes included in the 48 coding schemes. Virtually all the behavior coding schemes describe the basically structured behaviors in an interview and at least have the implicit or explicit goal of finding departures from the paradigmatic sequence in common. The behaviors are usually evaluated in terms of "adequate," "neutral" or "inadequate." However, depending on specific research questions, coding schemes often differ considerably from each other with respect to finer discriminations.

For example, a scheme may be developed to evaluate a specific type of interview (such as the Event History Calendar; see Belli et al. 2004).

Based upon the elements of the data collection process that in one way or another may affect the response obtained, we define four different foci of a coding scheme: interviewers, respondents, questions and the interaction. These elements are partly derived from Cannell and Oksenberg's (1988) distinction of goals of behavior coding. Studies can serve a meta-methodological goal (i.e., comparing different coding schemes or comparing behavior coding with other evaluation or pretest methods). However, the coding schemes in those meta-methodological studies can themselves always be classified according to the original focus, i.e., the element(s) they serve to pretest or evaluate. Schemes can also have multiple foci (e.g., Cannell, Fowler, and Marquis 1968; Belli et al. 2004).

In order to compare the different studies with respect to the aspects as discussed in the previous sections, and relate these aspects to the focus of the study, we will use a number of different categories that summarize the main characteristics of the coding scheme (see Table 6). We distinguished between three different aspects: the coding strategy, practical considerations in the coding procedure and the reliability of the scheme. Combining the two aspects of the coding strategy yields four different strategies: (a) selective coding at the Q-A sequence level (with no sequential information), which is often referred to as "conventional behavior coding," (b) selective coding at the exchange level, (c) selective coding at the utterance level, and (d) full coding with sequential information, which is often referred to as "interaction coding." The strategies (b) and (c) yield sequence information only at the exchange level. Therefore these two categories are integrated as one category. Additional aspects of a coding scheme are the number of actors involved (i.e., interviewer, respondent and possible third parties), the number of codes included and the mode of administration (face-to-face or telephone).

Table 6. Overview of aspects of comparison of behavior coding schemes

Aspect	Abr.	Specification
Coding strategy	SN	Selective coding at the Q-A sequence level, no sequence information
	SE	Selective coding with sequence information at the exchange level
	FS	Full coding with sequence information preserved
Coding procedure	L	Live coding
	Lr	Live coding, recording on cassette as backup
	Rc	Recorded tape coding
	Rt	Recorded transcript coding
	Rc/t	Recorded tape coding with transcripts as backup
Reliability procedure	K	Kappa
	KD	Kappa with unit of analysis *deviating* from level of coding
	P	Percentage
	PD	Percentage with unit of analysis *deviating* from level of coding
	C	Pearson correlation

5.1. The Interviewer as a Focus: Interviewer Monitoring Studies

As Cannell and Oksenberg (1988) point out, the results of interviewer monitoring studies can be used in terms of supervision ("enforcing rule following behavior") and evaluation (assessing the quality of particular studies, assessing overall staff performance, evaluating training methods, or exploring ways to improve training).

Especially many of the early behavior coding schemes are designed for the goal of monitoring interviewer performance (i.e., 14 of the 48 schemes compared). Table 7 shows that most coding schemes that were designed for interviewer monitoring use a selective coding scheme that does not preserve sequential information, and none of them uses a full coding scheme. Furthermore, many interview monitoring schemes include only interviewer behavior codes, such as Cannell, Lawson, and Hausser's (1975) scheme that served as a basis for many coding schemes (also for coding schemes with another focus, i.e., Morton-Williams 1979; Prüfer and Rexroth 1985; Sykes and Collins 1992). Their scheme included all the concepts and principles that were considered to be important targets in interviewer training. From this viewpoint the interviewer and respondent were considered individual actors that individually could produce errors.

5.1.1. Codes Included

Typically, interviewer monitoring schemes include the quality of question reading (distinguishing exact reading from reading with minor and/or major changes) and adherence to skip patterns. This *unconditional* scripted behavior mainly occurs before the respondent has spoken, therefore interviewers usually have direct control over it. Belli and Lepkowski (1996) conclude that "respondent behavior is more diagnostic of response accuracy than anything over which the interviewer has direct control" (Belli and Lepkowski 1996, p. 73). Therefore, it is very useful to also include codes that evaluate the interviewer's reaction to respondent behavior, i.e., *conditional* (un)scripted behavior. Furthermore, more than half of these coding schemes also include respondent behavior codes, which may be very relevant to evaluating interviewer behavior, e.g., to determining whether interviewers appropriately reacted to certain respondent behaviors.

5.1.2. Alternative Methods

Alternative assessments of interviewers' work (i.e., reviews of completed questionnaires, response distributions and progress monitoring of the number of interviews), although inexpensive and easily conducted, appear to reveal only a small part of inadequate interviewer performance (see Wilcox 1963, cited by Cannell and Oksenberg 1988). Such methods leave errors in the most important interviewer tasks (reading questions and probing) undetected. Direct observation (or listening-in) by a supervisor is usually subjective and unsystematic, but, as Cannell and Oksenberg state, "standardized coding of interviewer behavior provides an objective method for evaluating interviewer performance" (Cannell and Oksenberg 1988, p. 475).

5.2. The Questions as a Focus: Evaluating Questions

Another focus of a behavior coding scheme is to identify questions that cause problems for the interviewer or respondent, in order to pretest, evaluate or explore the effects of

Table 7. Coding schemes with interviewer behavior as focus

Scheme	Coding	Actors	Number of different codes		Procedure	Mode	Reliability procedure	Overall reliability
			I	R				
Cannell, Fowler, and Marquis (1968)	Selective, Non-sequential	Interviewer respondent	5	7	Live	Face-to-face	–	–
Cannell, Lawson, and Hausser (1975)	Selective, Exchanges	Interviewer	30	–	Recorded tape coding	Telephone	Kappa	.80–.92
Blair (1978)	Selective, Non-sequential	Interviewer respondent	39	11	Recorded tape coding	Face-to-face	Kappa	.74
Blair (1980)	Selective, Non-sequential	Interviewer	4	–	Recorded tape coding	Face-to-face	–	–
Bradburn and Sudman (1980)	Selective, Non-sequential	Interviewer respondent	4	2	Recorded tape coding	Face-to-face	Kappa	.52–.72
Mathiowetz and Cannell (1980)	Selective, Non-sequential	Interviewer	20	–	Live tape backup	Telephone	Percentage	.88
Prüfer and Rexroth (1985, Study 1)	Selective, Exchanges	Interviewer	35	–	Recorded tape coding	Face-to-face	–	–
Fowler and Mangione (1990)	Selective, Non-sequential	Interviewer	11	–	–	–	–	–
Shepherd and Vincent (1991, "compliance")	Selective, Non-sequential	Interviewer	16	–	Recorded tape coding	Telephone	–	–
Couper, Holland, and Groves (1992)	Selective, Non-sequential	Interviewer	16	–	Live	Telephone	–	–
Oksenberg, Cannell, and Blixt (1996)	Selective, Non-sequential	Interviewer respondent	14	7	Recorded tape coding		Kappa	.11–.90
Stanley (1996)	Selective, Non-sequential	Interviewer respondent	5	6	Recorded tape coding	Face-to-face	–	–
Brick, Tubbs et al. (1997b)	Selective, Non-sequential	Interviewer respondent	5	6	Recorded tape coding	Telephone	Percentage	.48–.68
Carton (1999)	Selective, Non-sequential	Interviewer respondent	41	12	Recorded tape coding	Face-to-face	–	–
Belli et al. (2004)	Selective, Exchanges	Interviewer respondent	25	17	Recorded tape coding, some transcripts	Face-to-face	Correlation	.42–1.0

Note: one overall reliability score or the minimum and maximum of all scores.

question wording. This focus has become more important since the first scheme of Morton-Williams (1979) and is the most frequently used focus of behavior coding schemes (i.e., it is to be found in 21 of the 48 schemes; see Table 8).

The rules that are the basis for these schemes and the codes that result from these schemes concern problematic interviewer behavior as well as problematic respondent behavior. For example, Morton-Williams (1979) gives nine criteria for adequate question performance. The categories of interviewer and respondent behavior she subsequently describes refer to the criteria on which a question might have failed.

5.2.1. Codes Included

Typically, coding schemes to evaluate questions include interviewer question reading codes and respondent codes that are assumed to indicate problems in question understanding (see Table 9 for an example of a complete scheme). However, these behavioral categories occur quite infrequently. As Schaeffer and Maynard (2002) also suggest, a number of other behavioral categories not typically included in behavior coding schemes (e.g., hesitations, reports, and feedback codes) may be much more effective in signaling problems with question wording, especially as compared to explicit requests for clarification, which respondents may avoid using, as a result of standardized interviewing practice.

Sources of problematic behaviors can often be found by means of comparison of the percentage of problematic behaviors and the specific question wording (Oksenberg, Cannell, and Kalton 1991). Close inspection of question wording may reveal why interviewers frequently change it, or why respondents frequently interrupt or give qualified answers. Furthermore, additional cues may be derived from answer distributions, information from coders and interviewers, and the transcripts, if available. Fowler's (1992) study illustrated the usefulness of behavior coding as a diagnostic tool that is also helpful to suggest revisions of question wording that improve the validity of data. Questions which were identified as problematic by means of behavior coding were redesigned. The alternative question wording not only yielded fewer instances of problematic behaviors, but also response distributions that were expected to be more accurate.

5.2.2. Alternative Methods

Behavior coding has often been judged as less effective in diagnosing problems with question wording than for instance cognitive interviewing (for a review, see Campanelli 1997). However, comparisons typically involve behavior coding in its usual implementation, i.e., in simple "selective" (see Section 2.1.5) coding schemes using only common codes such as those listed in Table 1.

In articles that compare behavior coding with other methods for their sensitivity of detecting problematic questions it is often recommended to use combinations of techniques, each yielding unique contributions (DeMaio et al. 1993; Presser and Blair 1994; Willis, Schechter, and Whitaker 1999; Hughes 2004). Furthermore, it is rather difficult or even useless to compare pretesting methods. Cognitive interviews have their own disadvantages, e.g., they can influence the question-answering process they seek to explore, because (especially concurrent) think-aloud instructions disturb this process. Moreover, respondents are not always able to spontaneously express their cognitive processes, especially when retrospective think aloud is applied (see Sudman, Bradburn, and Schwarz

Table 8. Coding schemes with the questions as focus

Scheme	Coding	Actors	Number of different codes		Procedure	Mode	Reliability procedure	Overall reliability
			I	R				
Morton-Williams (1979)	Selective, exchange levels	Interviewer respondent	14	17	Recorded tape coding	Face-to-face	–	–
Prüfer and Rexroth, 1985 (Study 2)	Full, sequential	Interviewer respondent	59	28	Recorded tape coding	Face-to-face	–	–
Sykes and Morton-Williams (1987, first study)	Selective, non-sequential	Interviewer respondent	1	5	Recorded tape coding	Face-to-face	–	–
Sykes and Morton-Williams (1987, second study)	Selective, non-sequential	Interviewer respondent	2	8	Recorded tape coding	Face-to-face	–	–
Oksenberg, Cannell and Kalton (1991)	Selective, non-sequential	Interviewer Respondent	3	7	–	–	Kappa	.60–.80
Gustavson-Miller, Herrman and Puskar (1991)	Selective, non-sequential	Interviewer respondent	9	6	Recorded tape coding	Face-to-face	Kappa	.55–.82
Esposito et al. (1992)	Selective, exchange levels	Interviewer respondent	6	7	Live	Telephone	Percentage deviating level	.87
Burgess and Patton (1993)	Selective, non-sequential	Respondent	–	5	Live	Telephone	–	–
DeMaio et al. (1993)	Selective, non-sequential	Interviewer respondent	6	6	Recorded tape coding	Telephone Face-to-face	–	–
Presser and Blair (1994)	Selective, non-sequential	Interviewer respondent	2	3	Live	Telephone	–	–
Cahalan et al. (1994)	Selective, non-sequential	Interviewer respondent	15	8	Recorded tape coding	Telephone	–	–
Blixt and Dykema (1995)	Selective, non-sequential	Respondent	–	5	Recorded tape coding	Face-to-face	Kappa	.65
Bates and Good (1996)	Selective, non-sequential	Interviewer respondent	4	5	Recorded transcript coding	Face-to-face	Percentage	.83
Dykema, Lepkowski, and Blixt (1997)	Selective, non-sequential	Interviewer respondent	4	6	Recorded tape coding	Face-to-face	–	–

Table 8. Continued

Scheme	Coding	Actors	Number of different codes		Procedure	Mode	Reliability procedure	Overall reliability
			I	R				
Hess, Singer, and Bushery (1997)	Selective, non-sequential	Interviewer respondent	5	8	Recorded tape coding	Telephone	Kappa	.55–.85
Lepkowski et al. (1998)	Selective, non-sequential	Interviewer respondent	6	13	Recorded tape coding	Face-to-face	Kappa	.18–.77
Snijkers (2002)	Selective, non-sequential	Respondent	–	7	Live	Telephone Face-to-face	–	–
Edwards et al. (2002)	Selective, non-sequential	Respondent	–	2	Live/ Recorded tape coding	Telephone	Kappa	.38
Van der Zouwen and Smit (2004)	Selective, non-sequential	Interviewer respondent	8	7	Recorded transcript coding	Face-to-face	Kappa deviating level	.76
Edwards et al. (2004)	Selective, non-sequential	Interviewer respondent	9	9	Recorded tape coding	Telephone	Kappa	0.0–1.0
Schaeffer and Dykema (2004)	Selective, exchange levels	Interviewer respondent	15	14	Recorded tape coding	Telephone	KD	.75 +

Note: Reliability one overall reliability score or the minimum and maximum of all scores.

Table 9. Oksenberg, Cannell and Kalton's scheme

Interviewer question reading codes	Respondent behavior codes
E Exact	1 Interruption with answer
S Slight change	2 Clarification
M Major change	3 Adequate answer
	4 Qualified answer
	5 Inadequate answer
	6 Don't know
	7 Refusal to answer

1996). With cognitive interviews, the chances of finding nonexistent problems are larger, whereas the chances of not finding existing problems are smaller than with behavior coding. However, behavior coding is often the only method that evaluates the interviewer objectively. Furthermore, behavior coding is often the only method that is quantitative and easy to replicate. Therefore, cognitive interviewing is ideally implemented in an operationalization phase, whereas behavior coding is ideally implemented in a pilot study of pretesting questions (Willis 2005).

5.3. The Respondent as a Focus

Monitoring respondent performance as a focus may seem odd at first sight, because a supervisor can hardly correct respondents. However, a researcher can monitor the behavior of respondents in survey interviews in order to identify and describe respondents difficult to interview. Four schemes that were (partly) designed to monitor respondent performance are summarized in Table 10.

Loosveldt (1997) used six respondent behavior categories as objective indicators of the respondent's cognitive and communicative skills. Gallagher, Fowler, and Roman (2004) tested the effects of training of aged respondents in their role, which appeared to be effective with respect to reducing the number of interruptions but not with respect to reducing interview length, response rates, or refusal rates.

5.3.1. Alternative Methods

Alternative assessments of response quality (i.e., item nonresponse and biases in response distributions), will reveal (like similar measurements to assess interviewer performance) only a small part of inadequate respondent behavior. Methods like interviewer debriefing or direct observation are also likely to be incomplete and subjective.

5.4. The Interaction as a Focus

Another goal of behavior coding studies can be to examine what the effects of specific behaviors will be on subsequent behaviors, or the interactional causes of specific behaviors. Hill and Lepkowski (1996) use the term "behavioral contagion" to indicate their goal to study how one instance of deviating behavior can lead to another.

Although all schemes that include evaluations of both interviewer and respondent behaviors may provide knowledge about the interaction, what is different about interaction

Table 10. Coding schemes with respondent behavior as a focus

Scheme	Coding	Actors	Number of different codes		Procedure	Mode	Reliability procedure	Overall reliability
			I	R				
Cannell et al. (1968)	Selective, non-sequential	Interviewer respondent	5	7	Live	Face-to-face	–	–
Loosveldt (1997)	Selective, non-sequential	Respondent	–	6	Recorded tape coding	Face-to-face	–	–
Belli et. al (2004)	Selective, exchange level	Interviewer respondent	25	17	Recorded tape coding, some transcripts	Telephone	Correlation	.42–1.0
Gallagher, Fowler, and Roman (2004)	Selective, exchange level	Interviewer respondent	15	9	Recorded tape coding	Telephone	–	–

schemes is the sequential information that is analyzed (i.e., in which order the behaviors occurred). As is shown in Table 11, all schemes include (some) sequential information. Furthermore, the table shows that all schemes include 20 or more codes, except for Hill and Lepkowski's (1996). The studies often have an explorative character (cf. Lepkowski, Siu, and Fisher 2000; Sykes and Collins 1992).

Marquis and Cannell (1969) conducted the first interactional study. As early as 1968, Cannell, Fowler, and Marquis reflected on a "reciprocal cue searching process" to be present in interviews. Their data led them to speculate about the existence of a process during the interview in which both interviewer and respondent are searching for cues about appropriate kinds of behavior (Cannell et al. 1968). Because the data in this study did not allow interactional analyses to prove these speculations to be right, in 1969 Marquis and Cannell used a revised coding scheme and coding procedure. They performed analyses on for instance the effects of directive and neutral probes on respondents giving adequate answers, or the probability that interviewer feedback follows specific categories of respondent behavior. Brenner (1982) recognized the importance of studying interactional processes (which he called "action-by-action" analysis) and was among the first who performed such analyses.

5.4.1. Codes Included

An important difference with respect to the codes included in schemes for interaction analysis as compared to other schemes, is that usually nonproblematic behaviors are also coded (i.e., reports, elaborations, perceptions, comments, etc.) in order to more fully describe the interaction (see Table 12 for an example). However, this requires a complex coding-scheme, and not all nonproblematic behaviors may be relevant. Using a summary code ("other behavior") can compensate for this problem. Although such a code will reduce the information available, it is possible to distinguish sequences with these summarizing codes from paradigmatic sequences. Therefore, it is always possible in a later stage to recode the summary codes into finer distinctions, if necessary.

5.4.2. Alternative Methods

Behavior coding suffers from the bias that it should be determined in advance what behaviors are relevant. Even full coding schemes do not always make fine discriminations, and may neglect distinctions that might be relevant afterwards. Therefore qualitative methods, such as conversation-analytic studies, may be useful. In that case transcripts are required, often using a detailed method of transcription according to conversation analysis conventions, as developed by Gail Jefferson (1983).

However, by using a full coding scheme with sequential information, where the original question wordings, and the entered responses are used, it is fairly possible that the data available to the researcher is close to the data available from transcripts of the interaction. In such a case, behavior coding may not only fulfill the requirements of availability of detailed data, but also enable a quantification of such data. Behavior coding enables a researcher to determine whether odd interactions are unusual incidents or evidence that data obtained by standardized interviews is untrustworthy. In this way, behavior coding may be helpful in resolving discussions between practitioners and critics of standardized interviewing (Maynard and Schaeffer 2002). Quantification is precisely what is lacking in qualitative data analysis, and this

Table 11. Coding schemes with the interaction as a focus

Scheme	Coding	Actors	Number of different codes		Procedure	Mode	Reliability procedure	Overall reliability
			I	R				
Marquis and Cannell (1969)	Full, sequential	Interviewer respondent	27	20	Recorde tape coding	F	–	–
Brenner (1982)	Full, sequential	Interviewer respondent	18	6	Recorded tape coding, some transcripts	F	–	–
Dijkstra et al. (1985) and Dijkstra (1983)	Full, sequential	Interviewer respondent third person	24	15	Recorded transcript coding	F	Kappa	.80
Loosveldt (1985)	Full, sequential	Interviewer respondent	95	79	Recorded tape coding	F	–	–
Shepherd and Vincent (1991,"interaction")	Selective, exchange level	Interviewer respondent	21	18	Recorded tape coding	T	–	–
Sykes and Collins (1992)	Selective, exchange level	Interviewer respondent	35	19	Recorded tape coding	F	Percentage, deviated level	.88
Smit (1995, simplified scheme)	Full, sequential	Interviewer respondent	10	10	Recorded transcript coding	F	Kappa	.72
Hill and Lepkowski (1996)	Selective, exchange level	Interviewer respondent	2	4	Live coding with tape	F	–	–
Dijkstra (1999, 2002)	Full, sequential	Interviewer respondent third person	± 139	± 60	Recorded transcript coding	–	Kappa	.78
Lepkowski, Siu, and Fisher (2000)	Selective, exchange level	Interviewer respondent	14	9	Recorded tape coding	F		–

Table 12. Brenner's coding scheme

Interviewer behavior codes		Respondent behavior codes
Question asked as required	Directive probing:	R answers adequately
Question asked with slight change	- based on R's information	R answers Don't know
Question significantly altered	- based on I's inference	R's information is inadequate
Question completely altered	Probing unrelated to task	R's information is irrelevant
Question asked directively	I repeats R's information	R gives feedback
Question omitted by mistake	I answers for R	R seeks clarification
Card omitted by mistake	I clarifies adequately	
Adequate probing	I gives feedback	
I repeats the question	I interrupts or closes Q-A sequence	
Leading probing		

is often used as a criticism of qualitative studies, as Houtkoop-Steenstra (2000) also notes.

6. Recommendations

The verbal behavior in a standardized interview yields a wealth of information that can be used for various goals. Because the behavior takes place in structured sequences of questions and answers, most coding schemes have many common elements.

In Table 13, the coding strategies and schemes that are recommended for different situations are listed. The choice of the coding strategy depends to an important extent on the focus and goal of the scheme. In the case of pretesting and monitoring (relevant parts of) the data collection it is important that quick results are available, in order to enable efficient processing of adaptations. This behavior coding takes place *prior to* or *during* actual data collection. Therefore schemes appropriate in this initial phase are limited to selective ones (with fewer than 15 codes) designed for frequency analysis. However, a scheme with more codes can be chosen when it is efficiently organized and high reliability scores have proven it to be feasible (e.g., Cannell, Lawson, and Hausser's scheme).

Performing behavior coding for evaluation or exploration (of relevant parts) of the data collection process can take place *after* actual data collection. In the case of evaluation quick results are not important, but detailed explanations of causes of problematic behaviors may not be relevant. Therefore selective coding schemes with a slightly higher number of codes (i.e., around 20) may be appropriate.

In the case of exploratory analyses of the interaction, detailed information is required, and full coding schemes with sequential information seem most appropriate. However, for practical application of such schemes, software like the Sequence Viewer program is necessary (see Dijkstra 2002).

Table 13. Recommended coding schemes for specific phase, goal and type of analysis

Focus	Type of study	Strategy	Type of analysis	Procedure	Examples of schemes
Interviewers	Monitoring	Selective	Frequency	Live	Cannell, Lawson, and Hausser (1975)
	Monitoring	Selective	Frequency	Tape	Brick et al. (1997b); Stanley (1996)
	Evaluation	Selective	Frequency	Tape	Oksenberg, Cannell, and Blixt (1996)
	Experiment	Selective	Frequency	Tape	Cannell Lawson, and Hausser (1975)
Questions	Pretest	Selective	Frequency	Live	Presser and Blair (1994)
	Pretest	Selective	Frequency	Tape	Oksenberg, Cannell, and Kalton (1991), DeMaio et al. (1993)
	Evaluation	Selective	Sequence (exchange)	Tape	Lepkowski, Siu, and Fisher (2000); Morton-Wiliams (1979)
	Exploration	Selective	Sequence (exchange)	Tape	Schaeffer and Dykema (2004)
	Experiment	Full	Sequence (utterances)	Transcript	Dijkstra (1999)
Respondents	Evaluation	Selective	Frequency	Tape	Gallagher (2004)
Interaction	Exploration	Full	Sequence (exchange)	Tape	Sykes and Collins (1992)
	Exploration	Full	Sequence (utterances)	Transcript	Dijkstra (1999)

7. Future Evaluations of Behavior Coding Schemes

In this article we have not empirically established differences between coding schemes. Forsyth, Rothgeb, and Willis (2004), following Willis et al. (1999), describe three general approaches to such methods evaluation (i.e., exploratory, confirmatory and reparatory research).

Exploratory and confirmatory research approaches compare methods with respect to how well they detect questionnaire problems. Behavior coding schemes can be compared for the difference (or subsequent confirmation or disconfirmation) in the information provided by different coding schemes when the same data are coded by different schemes (see Edwards et al. 2002 for a scheme with two categories that is compared with the scheme described in Oksenberg, Cannell, and Kalton 1991).

The reparatory approach, which, as Forsyth, Rothgeb, and Willis note, is rarely applied, compares methods for the effectiveness of suggested improvements. This research requires split-sample tests of questionnaires revised upon the basis of different coding schemes. Forsyth et al. (2004) followed this approach in a comparison of different pretest methods (i.e., expert review, questionnaire appraisal and cognitive interviews).

More research is needed on methods of pretesting the quality of questionnaires (Presser et al. 2004). As Fowler and Cannell state, "users of survey data lack information about the quality of the data collection process in general and the quality of the questions in particular. Behavior coding with its quantitative nature and its demonstrated relationship to key measures of data quality can provide indicators to readers on both subjects" (Fowler and Cannell 1996, p. 34).

Furthermore, although analysis of interviewer-respondent interactions will provide enough information about problems in survey interviews, computer-assisted questionnaire handling might also be an important element in the interaction. Interviewer-computer interaction might influence interviewer-respondent interaction and vice versa. This aspect has been largely ignored in behavior coding research. Schaeffer and Dykema (2004) anticipated this disturbing factor by including a coding option "CATI-problem" in their scheme. But, as Lepkowski et al. (1998) argue, behavior coding is not the appropriate method to study interviewer-computer interactions. In their study they compare behavior coding with usability evaluation, the latter being a method that can be used to study both interviewer-computer and interviewer-respondent interaction. Future evaluations of behavior coding schemes might therefore include such methods.

8. References

Bakeman, R. and Gottman, J.M. (1997). Observing Interaction: An Introduction to Sequential Analysis. Cambridge: University Press.

Bates, N. and Good, C. (1996). An Evaluation of the 1995 Test Census Integrated Coverage Measurement (Icm) Interview: Results from Behavior Coding. Paper presented at the Annual Meeting of the American Statistical Association. Chicago: U.S. Bureau of the Census.

Belli, R.F., Lee, E.H.L., Stafford, F.P., and Chou, C. (2004). Calendar Survey Methods: Association between Verbal Behaviors and Data Quality. Journal of Official Statistics, 20, 185–218.

Belli, R.F. and Lepkowski, J.M. (1996). Behavior of Survey Actors and the Accuracy of Response. Health Survey Research Methods: Conference Proceedings, DHMS Publication No. (PHS) 96-1013, 69–74.

Biemer, P., Herget, D., Morton, J., and Willis, G. (2000). The Feasibility of Monitoring Field Interview Performance Using Computer Audio Recorded Interviewing (CARI). Proceedings of the American Statistical Association, Section of Survey Research Methods, Alexandria, VA.

Blair, E. (1978). Nonprogrammed Speech Behaviors in a Household Survey. Unpublished doctoral dissertation, University of Illinois, Department of Business Administration.

Blair, E. (1980). Using Practice Interviews to Predict Interviewer Behaviors. Public Opinion Quarterly, 44, 257–260.

Blixt, S. and Dykema, J. (1995). Before the Pretest: Question Development Strategies. Proceedings of the American Statistical Association, Section of Survey Research Methods, Alexandria, VA, 1142–1147.

Bradburn, N.M. and Sudman, S. (1980). Improving Interview Method and Questionnaire Design. San Francisco: Jossey-Bass Publishers.

Brenner, M. (1982). Response-Effects of Role-Restricted Characteristics of the Interviewer. In Response Behaviour in the Survey-Interview, W. Dijkstra and J. Van der Zouwen (eds). London: Academic Press, 131–165.

Brick, J.M., Collins, M.A., Nolin, M.J., Davies, E., and Feibus, M.L. (1997a). Design, Data Collection, Monitoring, Interview Administration Time, and Data Editing in the 1993 National Household Education Survey. Technical Report, U.S. Department of Education. National Center for Education Statistics.

Brick, J.M., Tubbs, E., Collins, M.A., Nolin, M.J., Cantor, D., Levin, K., and Carnes, Y. (1997b). Telephone Coverage Bias and Recorded Interviews in the 1993 National Household Education Survey. Technical Report, National Center for Education Research.

Burgess, M.J. and Patton, D. (1993). Coding of Respondent Behaviour by Interviewers to Test Questionnaire Wording. Proceedings of the Americal Statistical Association, Section of Survey Research Methods, 392–397.

Cahalan, M., Mitchell, S., Gray, L., Chen, S., and Tsapogas, J. (1994). Recorded Interview Behavior Coding Study: National Survey of Recent College Graduates. Proceedings of the American Statistical Association, Section on Survey Research Methods.

Campanelli, P. (1997). Testing Survey Questions: New Directions in Cognitive Interviewing. Bulletin de Méthodologie Sociologique, 55, 5–17.

Cannell, C.F., Fowler, F.J., and Marquis, K.H. (1968). The Influence of Interviewer and Respondent Psychological and Behavioral Variables on the Reporting of Household Interviews. Vital and Health Statistics, Series 2, No. 26.

Cannell, C.F., Lawson, S.A., and Hausser, D.L. (1975). A Technique for Evaluating Interviewer Performance: A Manual for Coding and Analyzing Interviewer Behavior from Tape Recordings of Household Interviews. Technical Report, Survey Research Center of the Institute for Social Research, The University of Michigan.

Cannell, C.F. and Oksenberg, L. (1988). Observation of Behavior in Telephone Interviews. In Telephone Survey Methodology, R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls II, and J. Waksberg (eds). New York: Wiley, 475–495.

Carton, A. (1999). Een Interviewnetwerk: Uitwerking Van Een Evaluatieprocedure Voor Interviewers. Technical Report, Proefschrift Faculteit Sociale Wetenschappen. [In Dutch]

Couper, M.P., Holland, L., and Groves, R.M. (1992). Developing Systematic Procedures for Monitoring in a Centralized Telephone Facility. Journal of Official Statistics, 8, 63–76.

DeMaio, T.J., Mathiowetz, N.A., Rothgeb, J., Beach, M.E., and Durant, S. (1993). Protocol for Pretesting Demographic Surveys at the U.S. Census Bureau. Proceedings of the American Statistical Association, Section on Survey Research Methods.

Dijkstra, W. (1983). Beïnvloeding Van Antwoorden in Survey-Interviews. Academisch proefschrift Vrije Universiteit. Utrecht: Elinkwijk. [In Dutch]

Dijkstra, W. (1999). A New Method for Studying Verbal Interactions in Survey Interviews. Journal of Official Statistics, 15, 67–85.

Dijkstra, W. (2002). Transcribing, Coding, and Analyzing Verbal Interactions in Survey Interviews. In Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. Van der Zouwen (eds). New York: Wiley, 401–425.

Dijkstra, W. and Ongena, Y.P. (2002). Question-Answer Sequences in Survey-Interviews. Paper Presented at International Conference on Questionnaire Development, Evaluation and Testing Methods. Charleston, SC.

Dijkstra, W., Van der Veen, L., and Van der Zouwen, J. (1985). A Field Experiment on Interviewer-Respondent Interaction in the Research Interview; Uses and Approaches. In M. Brenner, J. Brown, and D. Cantor (eds). London: Academic Press, 37–63.

Dorsey, B.L., Rosemery, O.N., and Hayes, S.C. (1986). The Effects of Code Complexity and of Behavioral Frequency on Observer Accuracy and Interobserver Agreement. Behavioral Assessment, 8, 349–363.

Draisma, S. and Dijkstra, W. (2004). Response Latency and (Para)Linguistic Expressions as Indicators of Response Error. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: Wiley.

Dykema, J., Lepkowski, J.M., and Blixt, S. (1997). The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study. In Survey Measurement and Process Quality, L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: John Wiley and Sons, Inc.

Edwards, W.S., Fry, S., Zahnd, E., Lordi, N., and Willis, G. (2004). Behavior Coding Across Multiple Languages: The 2003 California Health Interview Survey as a Case Study. Proceedings of the American Statistical Association, Survey Research Methods Section.

Edwards, W.S., Narayanan, V., Fry, S., Catania, J.A. and Pollack, M. (2002). A Comparison of Two Behavior Coding Systems for Pretesting Questionnaires. Proceedings of the American Statistical Association, Survey Research Methods Section, 889–892.

Esposito, J.L., Rothgeb, J., Polivka, A.E., Hess, J., and Campanelli, P.C. (1992). Methodologies for Evaluating Survey Questions: Some Lessons from the Redesign of the Current Population Survey. Paper presented at International Conference on Social Science Methodology, Trento.

Forsyth, B., Rothgeb, J., and Willis, G. (2004). Does Pretesting Make a Difference? An Experimental Test. In Questionnaire Development, Evaluation and Testing Methods, S. Presser, J. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley.

Fowler, F.J. (1992). How Unclear Terms Affect Survey Data. Public Opinion Quarterly, 56, 218–231.

Fowler, F.J. and Cannell, C.F. (1996). Using Behavioral Coding to Identify Cognitive Problems with Survey Questions. In Answering Questions. Methodology for Determining Cognitive and Communicative Processes in Survey Research, N. Schwarz and S. Sudman (eds). San Francisco: Jossey-Bass, 15–36.

Fowler, F.J. and Mangione, T.W. (1990). Standardized Survey Interviewing; Minimizing Interviewer-Related Error. Newbury Park, CA: Sage.

Gallagher, P.M., Fowler, F.J., and Roman, A. (2004). Training Elderly Respondents: Does It Help? Paper presented at 59th Annual Conference of the American Association for Public Opinion Research. Phoenix, AZ.

Gustavson-Miller, L.A., Herrman, D.J., and Puskar, M.C. (1991). The Effects of Question Reading on Respondent Verbal Behavior. Technical Report, U.S. Bureau of Labor Statistics.

Hess, J., Rothgeb, J., and Zukerberg, A.L. (1997). Pretest Evaluation Report. Survey of Program Dynamics. Technical Report, Center for Survey Methods Research, U.S. Census Bureau.

Hill, D.H. and Lepkowski, J.M. (1996). Behavioral Contagion in the Health Field Survey. Paper presented at Health Survey Research Methods: Conference Proceedings.

Houtkoop-Steenstra, H. (2000). Interaction in the Standardized Survey Interview: the Living Questionnaire. Cambridge: Cambridge University Press.

Hughes, K.A. (2004). Comparing Pretesting Methods: Cognitive Interviews, Respondent Debriefing, and Behavior Coding. Technical Report, Statistical Research Division, U.S. Census Bureau.

Jefferson, G. (1983). Notes on a Possible Metric Which Provides for a "Standard Maximum" Silence of Approximately One Second in Conversation. In Till-Paper No. 42. Tilburg Papers in Language and Literature. Katholieke Universiteit Brabant.

Lepkowski, J.M., Couper, M.P., Hansen, S.E., Landers, W., McGonagle, K.A., and Schlegel, J. (1998). CAPI Instrument Evaluation: Behavior Coding, Trace Files, and Usability Testing. Proceedings of the American Statistical Association, Section of Survey Research Methods. Alexandria, VA.

Lepkowski, J.M., Siu, V., and Fisher, J. (2000). Event History Analysis of Interviewer and Respondent Survey Behavior. In Developments in Survey Methodology, A. Ferligoj and A. Mrvar (eds). Ljubljana: FDV, 3–20.

Loosveldt, G. (1985). De Effecten Van Een Interviewtraining Op De Kwaliteit Van Gegevens Bekomen Via Het Survey-Interview Technical Report, Proefschrift Katholieke Universiteit Leuven. [In Dutch]

Loosveldt, G. (1997). Interaction Characteristics of the Difficult to Interview Respondent. International Journal of Public Opinion Research, 9, 386–394.

Marquis, K.H. and Cannell, C.F. (1969). A Study of Interviewer-Respondent Interaction in the Urban Employment Survey. Ann Arbor: MI, Survey Research Center. University of Michigan.

Mathiowetz, N.A. (1999). Respondent Uncertainty as Indicator of Response Quality. International Journal of Public Opinion Research, 11, 289–296.

Mathiowetz, N.A. and Cannell, C.F. (1980). Coding Interviewer Behavior as a Method of Evaluating Performance. Proceedings of the American Statistical Association, Section of Survey Research Methods. Alexandria, VA, 525–528.

Maynard, D.W. and Schaeffer, N.C. (2002). Standardization and Its Discontents. In Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. Van der Zouwen (eds). New York: Wiley, 3–45.

Mazeland, H. (2003). Inleiding in De Conversatie Analyse. Bussum: Coutinho. [In Dutch]

Morton-Williams, J. (1979). The Use of "Verbal Interaction Coding" for Evaluating a Questionnaire. Quality and Quantity, 13, 59–75.

Oksenberg, L., Cannell, C.F., and Blixt, S. (1996). Analysis of Interviewer and Respondent Behavior in the Household Survey. National Medical Expenditure Survey Methods 7, Agency for Health Care Policy and Research. Technical Report, Public Health Service, U.S. Department of Health and Human Services.

Oksenberg, L., Cannell, C.F., and Kalton, G. (1991). New Strategies for Pretesting Survey Questions. Journal of Official Statistics, 7, 349–365.

Ongena, Y.P., Dijkstra, W., and Draisma, S. (2004). Conversational and Formal Questions in Survey Interviews. Paper presented at The Sixth International Conference on Logic and Methodology RC-33. Amsterdam.

Presser, S. and Blair, J. (1994). Survey Pretesting: Do Different Methods Produce Different Results? Sociological Methodology, 24, 73–194.

Presser, S., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Rothgeb, J., and Singer, E. (2004). Methods for Testing and Evaluating Survey Questionnaires. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, J. Rothgeb, and E. Singer (eds). New York: Wiley.

Prüfer, P. and Rexroth, M. (1985). Zur Anwendung Der Interaction-Coding-Technik ZUMA-Nachrichten, 17, 2–49. [In German]

Sacks, H., Schegloff, E., and Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking in Conversation. Language, 50, 696–735.

Schaeffer, N.C. and Dykema, J. (2004). Improving the Clarity of Closely Related Concepts: Distinguishing Legal and Physical Custody of Children. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley.

Schaeffer, N.C. and Maynard, D.W. (1996). From Paradigm to Prototype and Back Again: Interactive Aspects of Cognitive Processing in Standardized Survey Interviews. In Answering Questions. Methodology for Determining Cognitive and Communicative Processes in Survey Research, N. Schwarz and S. Sudman (eds). San Francisco: Jossey-Bass.

Schaeffer, N.C. and Maynard, D.W. (2002). Occasions for Intervention: Interactional Resources for Comprehension in Standardized Survey Interviews. In Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. Van der Zouwen (eds). New York: John Wiley and Sons, Inc.

Shepherd, J. and Vincent, C. (1991). Interviewer-Respondent Interactions in CATI Interviews. Proceedings of the Annual Research Conference. U.S. Bureau of the Census.

Smit, J.H. (1995). Suggestieve Vragen in Survey-Interviews. Vóórkomen, Oorzaken En Gevolgen. Amsterdam: Academisch proefschrift, Vrije Universiteit. [In Dutch]

Snijkers, G.J.M.E. (2002). Cognitive Laboratory Experiences. On Pre-Testing Computerised Questionnaires and Data Quality. Utrecht: Proefschrift Universiteit Utrecht.

Stanley, J.S. (1996). Standardizing Interviewer Behavior Based on the Results of Behavior Coding. Proceedings of the American Statistical Association, Survey Research Methods Section.

Sudman, S., Bradburn, N.M., and Schwarz, N. (1996). Thinking About Answers: The Application of Cognitive Processes to Survey Methodology. San Francisco: Jossey-Bass.

Sykes, W. and Collins, M. (1992). Anatomy of the Survey Interview. Journal of Official Statistics, 8, 277–291.

Sykes, W. and Morton-Williams, J. (1987). Evaluating Survey Questions. Journal of Official Statistics, 2, 191–207.

Van der Zouwen, J. (2002). Why Study Interaction in the Survey Interview? Response from a Survey Researcher. In Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. Van der Zouwen (eds). New York: Wiley, 47–65.

Van der Zouwen, J. and Dijkstra, W. (1998). Het Vraaggesprek Onderzocht. Wat Zegt Het Verloop Van De Interactie in Survey-Interviews over De Kwaliteit Van De Vraagformulering? Sociologische gids, 45, 387–403. [In Dutch]

Van der Zouwen, J. and Smit, J.H. (2004). Evaluating Survey Questions by Analyzing Patterns of Behavior Codes and Transcripts of Question-Answer Sequences: A Diagnostic Approach. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley.

Watson, D. (1982). The Actor and the Observer: How Are Their Perceptions of Causality Divergent? Psychological Bulletin, 92, 682–700.

Wilcox, K. (1963). Comparison of Three Methods for Collection of Morbidity Data by Household Survey. Ph.D. Dissertation, the University of Michigan, School of Public Health.

Willis, G. (2005). Cognitive Interviewing. A Tool for Improving Questionnaire Design. Thousand Oaks: Sage.

Willis, G., Schechter, S., and Whitaker, K. (1999). A Comparison of Cognitive Interviewing, Expert Review and Behavior Coding: What Do They Tell Us? Proceedings of the American Statistical Association, Survey Research Methods Section, 28–37.

Received September 2004
Revised November 2005